1 Multicollinearity

- ***Multicollinearity*** refers to **correlation** between **predictor** variables.
- For insight on the effects of *multicollinearity* on a regression analysis, consider the **two extremes** involving a model with **two predictors**:

  1. $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are **uncorrelated**. In this case,
     - ▷ The estimated coefficient for one predictor, b_2 say, will be the same no matter whether or not X_1 is included in the model.
     - ▷ The extra sum of squares $\text{SSR}(X_2|X_1)$ will be the same as the usual regression sum of squares $\text{SSR}(X_2)$ obtained from a simple linear regression of Y on X_2 alone.
     - ▷ Thus when both X_1 and X_2 are included in model, the marginal contribution of each in explaining Y variation is the same as it would be if that predictor was used by itself in a simple linear regression analysis.
  2. $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are **perfectly correlated**. In this case,
     - ▷ Infinitely many response planes will fit the data equally well (i.e. will result in the same set of fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$).
     - ▷ This means that there will be infinitely many choices of b_1 and b_2, all of which produce regression planes that fit the data equally well.
     - ▷ Although a model can be fitted, statistical software will often give an warning message because the matrix $\mathbf{X}^T\mathbf{X}$ will be singular, and therefore non-invertible.
     - ▷ You could do just as well (in terms of explaining Y variation) by performing a simple linear regression with just one of the two predictors – the other predictor is redundant in the sense that it doesn't add any new information to the model.

- In between the two extremes, multicollinearity may result in the following problems:

  1. Some of the estimated standard errors $s\{b_1\}, s\{b_2\}, \ldots, s\{b_{p-1}\}$ may be **very large**. As a consequence, individual predictors **may not be statistically significant** (according to the t tests), even though the overall model F test result may be **highly significant**. Large standard errors are is therefore one way to diagnose whether multicollinearity exists.
  2. The marginal effect of a predictor added to the model (e.g. as measured by an extra sum of squares) can be **different** depending on which other predictors are already in the model.

3. An estimated coefficient $\boldsymbol{b_k}$ may **differ** depending on which other predictors are included in the model.
4. The results of the t test for a coefficient β_k may **differ** depending on which other predictors are included in the model.

- If the *only* objective is to make predictions $\hat{Y}_h$ or draw inferences about $\mathrm{E}(Y_h)$ within the range of the observed X data, multicollinearity is generally not a problem. But if the objective is to draw inferences about individual coefficients $\beta_1, \beta_2, \ldots, \beta_{p-1}$, multicollinearity can be a problem.
- We'll see how to assess multicollinearity and remedy problems arising from it later.

1.1 General Linear Models

- Simple linear and multiple regression models are special cases of a class of models called *general linear models*. In ***general linear models***:
  - ▷ The predictors can include **categorical** variables coded by ***indicator*** variables. For example:
$$X_i = \begin{cases} 0 & \text{if the } i\text{th individual is male} \\ 1 & \text{if the } i\text{th individual is female} \end{cases}$$
(***Analysis of variance***, or ***ANOVA***, models are *general linear models* in which **all** of the predictors are categorical).
  - ▷ The predictors can include **both** numerical **and** categorical variables. (***Analysis of covariance***, or ***ANCOVA***, models are *general linear models* that include **both** numerical **and** categorical predictors).
  - ▷ ***Transformations*** of the predictors are allowed in the model. For example *polynomial regression* regression models such as
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{1}$$
or models such as
$$Y_i = \beta_0 + \beta_1 \log(X_i) + \epsilon_i \,. \tag{2}$$
  - ▷ ***Interactions*** between the predictors are allowed in the model, for example
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i \,. \tag{3}$$

  Note that the mean responses in (1) and (2) are **curves**, and the mean response in (3) is a **curved surface**.
- All that's required of a model to be a **general linear model** is that, for fixed values of the predictors $X_1, X_2, \ldots, X_k$, the mean response $E(Y)$ is a **linear combination** of the **parameters** $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$.

2 Polynomial Regression

- ***Polynomial regression models*** are a special case of *general linear models.*

2.1 Polynomial Regression with One Predictor

- A ***quadratic*** (or ***second-order***) regression model with one predictor is

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \,. \tag{4}$$

- A ***cubic*** (or ***third-order***) regression model with one predictor is

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i \,. \tag{5}$$

- Higher order polynomial regression models can also be fit. Polynomial regression models are commonly used to approximate a nonlinear, but unknown, relationship between a response variable Y and predictor X.

- The **design matrix** for the **quadratic** regression model (4) is

$$\mathbf{X} \;=\; \begin{bmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{bmatrix}$$

  The **design matrix** for the **cubic** regression model (5) is

$$\mathbf{X} \;=\; \begin{bmatrix} 1 & X_1 & X_1^2 & X_1^3 \\ 1 & X_2 & X_2^2 & X_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 & X_n^3 \end{bmatrix}$$

- In either case, the **vector $\boldsymbol{b}$** of least squares **estimates $\boldsymbol{b_0, b_1, \ldots, b_{p-1}}$** of the model parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ is obtained by

> **Least Squares Estimates of $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_{p-1}}$ (Matrix Approach)**: The vector of estimated coefficients $\mathbf{b}$ is obtained by:
>
> $$\mathbf{b} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{6}$$

- The ***fitted polynomial regression model*** is

Fitted Polynomial Regression Model:

$$\hat{Y} \;=\; b_0 + b_1 X + b_2 X^2 + \cdots + b_{p-1} X^{p-1}.$$

- The variance-covariance matrix of $\mathbf{b}$ is

$$\sigma^2\{\mathbf{b}\} \;=\; \sigma^2 \cdot (\mathbf{X}^T\mathbf{X})^{-1}$$

and the (estimated) variance-covariance matrix $\boldsymbol{s^2\{b\}}$ is obtained by replacing σ^2 by MSE. Thus the **(estimated) standard errors** of $b_0, b_1, \ldots, b_{p-1}$ reported by statistical software are:

(Estimated) standard errors: The (estimated) standard errors $\boldsymbol{s\{b_0\}}$, $\boldsymbol{s\{b_1\}}$, $\boldsymbol{\ldots, s\{b_{p-1}\}}$ of $b_0, b_1, \ldots, b_{p-1}$ are the square roots of the diagonal elements of the matrix

$$s^2\{\mathbf{b}\} \;=\; \text{MSE} \cdot (\mathbf{X}^T\mathbf{X})^{-1}\,. \tag{7}$$

2.2 Polynomial Regression on Centered Variables

- **Multicollinearity** can be a problem in polynomial regression (e.g. X and X^2 are correlated, so the matrix $\mathbf{X}^T\mathbf{X}$ may be nearly singular).
- To alleviate the problem, we can **center** the predictor variable X, i.e. for each observed value X_i calculate

$$x_i \;=\; X_i - \bar{X}\,,$$

and then fit the model to the **centered** predictor, e.g. fit

$$Y_i \;=\; \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i\,.$$

Centering the predictor before fitting the polynomial regression model *reduces problems arising from multicollinearity.*

- Once the regression model has been fit using the **centered** variable x, it's possible to re-express the fitted model in terms of the uncentered variable X. See the textbook.

2.3 Deciding Upon the Appropriate Order of the Polynomial

- To decide upon the order of a polynomial model we can use the *partial F test approach* to decide whether the higher order terms can be dropped from the model.

  For example, to decide whether the term X^3 can be dropped from a cubic polynomial model, the test statistic is

$$F = \frac{\text{MSR}(X^3|X, X^2)}{\text{MSE}(X, X^2, X^3)}.$$

  If a **higher order** term (e.g. X^3) is retained in the model, then all **lower order** terms (e.g. X and X^2) should also be **retained**.

2.4 Polynomial Regression with Two or More Predictors

- Polynomial models can be fit with **more than one** predictor variable.

  For example, a ***second-order*** regression model with **two predictors**, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \beta_4 X_{i2}^2 + +\beta_5 X_{i1} X_{i2} + \epsilon_i$$

  and the design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{11}^2 & X_{12}^2 & X_{11}X_{12} \\ 1 & X_{21} & X_{22} & X_{21}^2 & X_{22}^2 & X_{21}X_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n1}^2 & X_{n2}^2 & X_{n1}X_{n2} \end{bmatrix}$$

- Estimates $b_0, b_1, \ldots, b_{p-1}$ of the model parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are obtained through (6), and standard errors of the estimators are obtained from the diagonals of the matrix (7).

- Higher order models can be fit, and more than two predictors can be incorporated. As was the case in polynomial regression with one predictor, it's best to center the predictors before fitting the model to alleviate problems from multicollinearity.